Structure Selection for Convolutive Non-negative Matrix Factorization Using Normalized Maximum Likelihood Coding

Atsushi Suzuki
Graduate School of
Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo
113-8654 Japan
atsushi_suzuki@mist.i.u-tokyo.ac.jp

Kohei Miyaguchi
Graduate School of
Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo
113-8654 Japan
kohei_miyaguchi@mist.i.u-tokyo.ac.jp

Kenji Yamanishi
Graduate School of
Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo
113-8654 Japan
yamanishi@mist.i.u-tokyo.ac.jp

***Abstract*—Convolutive non-negative matrix factorization (CNMF) is a promising method for extracting features from sequential multivariate data. Conventional algorithms for CNMF require that the structure, or the number of bases for expressing the data, be specified in advance. We are concerned with the issue of how we can select the best structure of CNMF from given data. We first introduce a framework of probabilistic modeling of CNMF and reduce this issue to statistical model selection. The problem is here that conventional model selection criteria such as AIC, BIC, MDL cannot straightforwardly be applied since the probabilistic model for CNMF is irregular in the sense that parameters are not uniquely identifiable. We overcome this problem to propose a novel criterion for best structure selection for CNMF. The key idea is to apply the technique of latent variable completion in combination with normalized maximum likelihood coding criterion under the minimum description length principle. We empirically demonstrate the effectiveness of our method using artificial and real data sets.**

I. Introduction

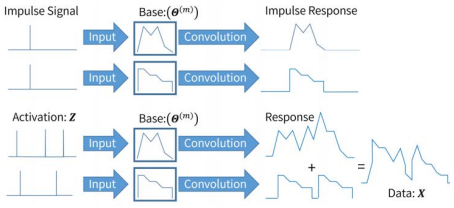

Fig. 1. The intuitive interpretation of convolution process from bases and their activations to data. In this example, the number of bases K is 2.

A. Motivation and Purpose

In this paper we are concerned with the issue of how we can estimate the best structure for *convolutive non-negative matrix factorization* (CNMF)[11]. The task of CNMF is to decompose matrix $\boldsymbol{X} \in \mathbb{R}^{\Omega \times T}$ into convolution of a number of bases $\boldsymbol{\Theta} \stackrel{\text{def}}{=} \left(\boldsymbol{\Theta}^{(m)}\right)_{m=0}^{M-1} = \boldsymbol{\Theta}^{(0)}, \boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(M-1)}$ ($\boldsymbol{\Theta}^{(m)} \in \mathbb{R}^{\Omega \times K}$ for $m = 0, 1, \ldots, M-1$) and their activations $\boldsymbol{Z} \in \mathbb{R}^{K \times T}$ as $\boldsymbol{X} \approx \sum_{m=0}^{M-1} \boldsymbol{\Theta}^{(m)} \overset{m\rightarrow}{\boldsymbol{Z}}$. We let $\overset{m\rightarrow}{\cdot}$ ($\overset{\leftarrow m}{\cdot}$) denote a right (left) column shift operator that moves its argument m places to the right (left); as each column is shifted off to the right (left), the leftmost (rightmost) columns are filled by zeros. Here $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$ consist of K latent factors, where we denote the kth component of $\boldsymbol{\Theta}^{(m)}$ as $\boldsymbol{\theta}_k^{(m)}$ ($m = 0, 1, \ldots, M-1$) and the kth row of $\boldsymbol{Z}$ as $\boldsymbol{z}_k^{\mathrm{T}}$.

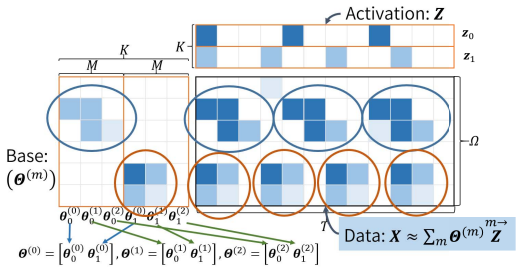

Fig. 2. The intuitive interpretation of extracting time sequential features from data. In this example, $\Omega = 6$, $K = 2$, $T = 12$, and $M = 3$, respectively. Given data $\boldsymbol{X}$, CNMF gets bases $\boldsymbol{\Theta}$ and their activation matrices $\boldsymbol{Z}$.

An example of the convolution process is illustrated in Fig. 1. An impulse response is generated as a convolution of a number of bases when a single impulse signal is input. For an individual system, a sequence of responses is generated as a convolution of a number of bases when a sequence of multi impulses is input as an activation. For the overall system, the data is observed as a linear combination of the individual systems' responses. CNMF is a process of decomposing the data into the convolution of bases and activation.

We may conduct CNMF for a data sequence to extract K underlying time sequential features from it. We may think of $\left\{\boldsymbol{\theta}_k^{(0)}, \boldsymbol{\theta}_k^{(1)}, \ldots, \boldsymbol{\theta}_k^{(M-1)}\right\}$ as the kth sequential features. We refer to K as *the number of bases* of CNMF. An example of extracting 2 features with CNMF is illustrated in Fig. 2.

The conventional *non-negative matrix factorization* (NMF) can be thought of as a special case of CNMF with $M = 1$. CNMF with $M \geq 2$ is more appropriate for analyzing sequential data than NMF in the case where time sequential features may underlie the given data. This is because CNMF views the original data as the sum of the K latent factors and time-dependent natures may be represented through the differences of $\left(\boldsymbol{\Theta}^{(m)}\right)_{m=0}^{M-1}$.

The most critical issue with CNMF is how we can determine the number of bases for CNMF. All of the conventional algorithms for CNMF have been designed under the assumption that the number of bases is given in an ad hoc way. However, it is reasonable to select the best number on the basis of given data. The purpose of this paper is to propose a novel method for selecting the best structure of CNMF from the view of statistical model selection.

B. Novelty and Significance of This Paper

We summarize our contributions as follows:

a) Proposing algorithm for selecting the best structure of CNMF: Our theoretical contributions are listed as follows:
a-1) Developing a new framework for selecting the best structure of CNMF:

We first introduce a framework of probabilistic modeling of CNMF. This is necessary for taking a statistical approach into the structure selection for CNMF. Our modeling can be thought of as a natural extension of Virtanen et al.'s latent variable model for NMF [12]. Within our framework we derive an efficient algorithm for estimating both parameters and latent variables using the *auxiliary function method*.

a-2) Novel base selection criterion using normalized maximum likelihood coding:

Conventional statistical model selection criteria such as AIC[1], BIC[9], or MDL[7] cannot straightforwardly be applied into our framework as above. This is because our model CNMF is specified by latent variables, which makes the model *irregular* in the sense that there are some parameter values which cannot uniquely be identified. Meanwhile most of statistical model selection criteria are designed under the condition that the central limit theorem holds. This condition is not fulfilled when the model is irregular.

To overcome this difficulty, we first apply the technique of *latent variable completion* (LVC) [3], [4] into our probabilistic model to obtain a regular model. In LVC the likelihood of the joint distribution of data and latent variables is calculated where the values of latent variables are estimated from data. We then select the number of bases in the resulting regular model on the basis of the *minimum description length* (MDL) principle. That is, we select the number of bases so that the total code-length required for encoding latent variables and the underlying regular model as well as data is minimum. We employ the MDL principle because it 1) has consistency [8], and 2) is guaranteed rapid convergence in the framework of probably approximately correct (PAC) learning [16].

We derive a novel MDL criterion formula for CNMF structure selection by employing the *normalized maximum likelihood* (NML) coding with LVC. Note that it is critical what kind of coding is used for calculating code-lengths. The reason why we employ the NML coding is that it is optimal in the sense that it attains Shtarkov's minimax criteria [10] . To the best of the authors' knowledge, this is the first work on selecting the number of bases as well as the parameter values for CNMF.

b) Empirical demonstration of effectiveness of CNMF in data mining: We empirically demonstrate the effectiveness of our method through artificial data and real data.

For artificial data sets, we investigate how well our method can estimate the number of bases as the temporal length of data increases. We show that our method significantly outperforms the state of arts; AIC, BIC with LVC. As for experiments with real data sets, we employ price fluctuation data set. We show that our proposed method is able to estimate the most reasonable number of bases, and eventually to discover important features underlying the time-varying real data.

C. Related Works

Since the original framework of CNMF was developed by [11], any probabilistic modeling has not been made for it. There exist a number of works on probabilistic modeling for NMF (e.g. Vitranen et al. [12], Cemgil [2], Ito et al. [4]). Our modeling can be thought of as a natural extension of Virtanen's one into the case where the convolution of NMFs is made.

The base selection problem for CNMF can be reduced to statistical model selection once the probabilistic framework is build. There exist a number of effective statistical model selection methods including Akaike's information criterion (AIC) [1], the Bayesian information criterion [9], the minimum description length (MDL) [7] criterion, and the minimum message length (MML)[13]. The problem is that these criteria cannot straightforwardly be applied to the probabilistic model of CNMFs, because it is irregular as previously discussed.

There exist a number of recent works which tackle with this problem. Watanabe derived the widely-applicable Akaike's information criterion (WAIC) [14] as an unbiased estimator of the generalized loss, and the widely-applicable Bayesian information criterion (WBIC) [15] as an asymptotically consistent estimator of the marginal log-likelihood, for irregular models respectively. However, these criteria work only when a datum is independently generated. Hence we cannot apply them to the CNMF structure selection because a datum at a moment depends on that at different moments in CNMF.

We solve this problem by employing a technique of latent variable completion (LVC) in combination with the normalized maximum likelihood (NML) code-length criterion in the scenario of the MDL principle. LVC has been applied to model selection for latent variable models. Ito et al.[4] proposed a rank selection method for NMF by employing LVC with NML where the independence assumption is made. It cannot straightforwardly be applied to the structure selection for CNMF because the convolutive structure of CNMF makes the computation of NML more difficult.

II. Convolutive Non-negative Matrix Factorization

A. Conventional Formulation

Convolutive non-negative matrix factorisation (CNMF) is is formulated as follows. Given a non-negative matrix $\boldsymbol{X} \in \mathbb{R}_+^{\Omega \times T}$, the goal is to approximate $\boldsymbol{X}$ as a convolution of bases $\boldsymbol{\Theta}^{(m)}$ and their activity $\boldsymbol{Z}$,

$$\boldsymbol{X} \approx \sum_{m=0}^{M-1} \boldsymbol{\Theta}^{(m)} \overset{m\rightarrow}{\boldsymbol{Z}},$$

where $\boldsymbol{\Theta}^{(m)} \in \mathbb{R}_+^{\Omega \times K} (m = 0, 1, \ldots, M-1)$ and $\boldsymbol{Z} \in \mathbb{R}_+^{K \times T}$. It is element-wisely written as follows:

$$x_{\omega t} \approx \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} \theta_{\omega k}^{m} z_{k(t-m)}.$$

Let $\boldsymbol{\Lambda} \in \mathbb{R}_+^{\Omega \times T}$ denote the approximated value:

$$\boldsymbol{\Lambda} \overset{\text{def}}{=} \sum_{m=0}^{M-1} \boldsymbol{\Theta}^{(m)} \overset{m\rightarrow}{\boldsymbol{Z}}.$$

It is element-wisely written as follows:

$$\lambda_{\omega t} = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} \theta_{\omega k}^{(m)} z_{k(t-m)}.$$

B. Approximation Criterion

Let $f_{\omega t} \in \{0, 1\}$ denote the presence of the datum $f_{\omega t} \in \{0, 1\}$ ($f_{\omega t} = 0$ if the datum is missed and $f_{\omega t} = 1$ otherwise) and set $\boldsymbol{F} \overset{\text{def}}{=} (f_{\omega t})$.

We employ the Kullback-Leibler divergence as a criterion or evaluating the goodness of approximation between two matrices $\boldsymbol{X}$ and $\boldsymbol{\Lambda}$ as follows:

$$D_{\mathrm{KL}} (\boldsymbol{X} \| \boldsymbol{\Lambda}) \overset{\text{def}}{=} \sum_{\omega=0}^{\Omega-1} \sum_{t=0}^{T-1} \left(f_{\omega t} x_{\omega t} \ln \frac{x_{\omega t}}{\lambda_{\omega t}} - f_{\omega t} x_{\omega t} + f_{\omega t} \lambda_{\omega t} \right).$$

The values of $\boldsymbol{\Lambda}$ in CNMF are estimated by minimizing this criterion for given X. This is the conventional formulation of CNMF introduced by Smaragdis [11]. Note that we cannot apply statistical model selection criteria to CNMF of this formulation, since no probabilistic assumption is made there.

III. Probablistic Modeling

A. Probabilistic Formulation

In order to consider CNMF from the view of statistical model selection, we introduce a framework of its probabilistic modeling. This is based on the Bayesian approach to *non-negative matrix factorization* (NMF) developed by [12]. We interpret the bases $\boldsymbol{\Theta}$ as the *parameters* of the probabilistic model, and re-interpret the activities $\boldsymbol{Z}$ as a *latent variable* generated by

$$z_{kt} \sim \mathrm{Gamma} (z_{kt}; a, b).$$

We further introduce an intermediate latent variable $\boldsymbol{S} \in \mathbb{R}_+^{\Omega \times K \times M \times T}$, which is assumed to be generated as follows:

$$s_{\omega k t m} \sim \mathrm{Poisson} \left(s_{\omega k t m}; \theta_{\omega k}^{(m)} z_{k(t-m)} \right).$$

It is specified by a parameter vector $\boldsymbol{\Theta}$ and is conditioned on $\boldsymbol{Z}$. We further assume that the data $\boldsymbol{X}$ is generated as follows:

$$x_{\omega t} \sim \delta \left(x_{\omega t} - \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} s_{\omega k t m} \right).$$

The goal of CNMF is to estimate $(\boldsymbol{Z}, \boldsymbol{\Theta})$ so that the joint probability of $\boldsymbol{X}$ and $\boldsymbol{Z}$ parametrized by $\boldsymbol{\Theta}$ is maximized with respect to $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$. That is, the problem is reduced to the maximization problem:

$$\underset{\boldsymbol{Z}, \boldsymbol{\Theta}}{\operatorname{argmax}} p (\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\Theta}).$$

This is equivalent with the problem of minimizing the following loss function:

$$\begin{aligned} & L (\boldsymbol{X}; \boldsymbol{Z}, \boldsymbol{\Theta}) \\ & \overset{\text{def}}{=} \sum_{\omega=0}^{\Omega-1} \sum_{t=0}^{T-1} \left(f_{\omega t} x_{\omega t} \ln \frac{x_{\omega t}}{\lambda_{\omega t}} - f_{\omega t} x_{\omega t} + f_{\omega t} \lambda_{\omega t} \right) + R (\boldsymbol{Z}), \end{aligned} \tag{1}$$

where $R (\boldsymbol{Z}) \overset{\text{def}}{=} \frac{z_{kt}}{a} + \ln \Gamma(b) a^b - \sum_{k=0}^{K-1} \sum_{t=0}^{T-1} (b-1) \ln z_{kt}$. Here $R (\boldsymbol{Z})$ has a role of regularizing the scale of $\boldsymbol{Z}$. Without this term $(\boldsymbol{Z}, \boldsymbol{\Theta})$ would give the same probability as $\left(c\boldsymbol{Z}, \frac{1}{c}\boldsymbol{\Theta}\right)$.

B. Parameter Estimation Algorithm

The minimization problem for the loss function (1) is intractable because its partial differential coefficient with respect to an element includes other elements. We take the auxiliary function approach to this problem. The effectiveness and stability of this approach have been demonstrated in many of previous works; e.g. [6], [5] etc. We propose an efficient optimization method for this loss function on the bases of the auxiliary function approach. Minimizing (1) is equivalent to minimizing $L^0 (\boldsymbol{X}; \boldsymbol{Z}, \boldsymbol{\Theta})$ defined as

$$\begin{aligned} L^0 (\boldsymbol{X}; \boldsymbol{Z}, \boldsymbol{\Theta}) \overset{\text{def}}{=} & \sum_{\omega=0}^{\Omega-1} \sum_{t=0}^{T-1} (-f_{\omega t} x_{\omega t} \ln \lambda_{\omega t} + f_{\omega t} \lambda_{\omega t}) \\ & + \sum_{k=0}^{K-1} \sum_{t=0}^{T-1} \left[-(b-1) \ln z_{kt} + \frac{z_{kt}}{a} \right] \end{aligned}$$

We introduce an *auxiliary function* $L^+ (\boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\mu}; \boldsymbol{X})$ such that

$$\begin{aligned} & L^+ (\boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\mu}; \boldsymbol{X}) \\ & \overset{\text{def}}{=} \sum_{\omega=0}^{\Omega-1} \sum_{t=0}^{T-1} \left(-f_{\omega t} x_{\omega t} \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} \mu_{\omega k t m} \ln \frac{\theta_{\omega k}^{(m)} z_{k(t-m)}}{\mu_{\omega k t m}} \right. \\ & \left. + f_{\omega t} \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} \theta_{\omega k}^{(m)} z_{k(t-m)} \right) - \sum_{k=0}^{K-1} \sum_{t=0}^{T-1} \left[(b-1) \ln z_{kt} - \frac{z_{kt}}{a} \right], \end{aligned}$$

where $\boldsymbol{\mu} = (\mu_{\omega k t m})$ is an auxiliary variable. It is proved that by Jensen's inequality, we have

$$L^0 (\boldsymbol{X}; \boldsymbol{Z}, \boldsymbol{\Theta}) = \min_{\boldsymbol{\mu}} L^+ (\boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\mu}; \boldsymbol{X}).$$

We show an outline of our matrix-form algorithm for parameter estimation in Algorithm 1. Here $\odot$ and $\oslash$ denote

Algorithm 1 CNMF Algorithm Based on Auxiliary Function Approach (Matrix-form)

repeat

$$\boldsymbol{Z} \leftarrow \left[(b-1)\mathbf{1}_{K\times T} + \boldsymbol{Z} \odot \left(\sum_{m=0}^{M-1} \boldsymbol{\Theta}^{(m)\mathrm{T}} \left[\boldsymbol{F} \odot \overleftarrow{\boldsymbol{X}}^{m} \oslash \boldsymbol{\Lambda}\right]\right)\right] \oslash \left[\frac{1}{a}\mathbf{1}_{K\times T} + \sum_{m=0}^{\min\{M-1,t\}} \boldsymbol{\Theta}^{(m)\mathrm{T}} \overleftarrow{\boldsymbol{F}}^{m}\right]$$

for $m = 0, 1, \ldots, M-1$ **do**

$$\boldsymbol{\Theta}^{(m)} \leftarrow \left[\boldsymbol{\Theta}^{(m)} \odot \left(\boldsymbol{F} \odot \boldsymbol{X} \oslash \boldsymbol{\Lambda}\right) \overrightarrow{\boldsymbol{Z}}^{m\mathrm{T}}\right] \oslash \left[\boldsymbol{F} \overrightarrow{\boldsymbol{Z}}^{m}\right]$$

end for

until it converges

elementwise multiplication and division, respectively. Note that if $b \geq 1$, the parameters remain positive during update. Because this algorithm is derived on the auxiliary function method, it is guaranteed that this algorithm monotonously decreases the original loss function $L(\boldsymbol{X}; \boldsymbol{Z}, \boldsymbol{\Theta})$. The computational cost in a step in this algorithm is $O(\Omega KTM)$, while it is $O(\Omega KT)$ for NMF (CNMF with $M = 1$).

IV. Structure Selection

A. Irregularity Problem

Any of parameter estimation algorithms for CNMF requires that the number of bases be specified in advance. To specify the number of bases, statistical model selection would be applied to our probabilistic framework for CNMF, but we may suffer from the *irregularity problem*, as shown below.

We say that the model is *irregular* when the map from a parameter to a probability distribution is not one-to-one. In CNMF case, there exists $(\boldsymbol{Z}', \boldsymbol{\Theta}') \neq (\boldsymbol{Z}, \boldsymbol{\Theta})$ such that $\sum_{m=0}^{M-1} \boldsymbol{\Theta}'^{(m)} \boldsymbol{Z}' = \sum_{m=0}^{M-1} \boldsymbol{\Theta}^{(m)} \boldsymbol{Z}$, which implies that parameters cannot be uniquely identified from $\boldsymbol{X}$. Thus CNMF is an irregular model.

When the model is irregular, the central limit theorem for the maximum likelihood estimator does not hold. Thus, conventional statistical model selection criteria such as AIC [1], BIC [9], and the conventional form of MDL principle [7] cannot straightforwardly be applied.

B. Normalized Maximum Likelihood Codes

To overcome the irregularity problem, we propose a new method for structure selection for CNMF: combining the latent variable completion with normalized maximum likelihood code-length criterion under the MDL principle.

According to the MDL principle, we select the best model so that the total code-length required for encoding both the data and model is minimum. Specifically we employ the *normalized maximum likelihood* (NML) coding in the calculation of code-lengths. We calculate the NML code-lengths for $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}$ relative to the *complete variable model* rather than the marginal distribution of $\boldsymbol{X}$. The complete variable model is the joint distribution of data $\boldsymbol{X}$ and latent variables $\boldsymbol{Z}, \boldsymbol{S}$. It is no longer an irregular model, since the parameter is uniquely identifiable once the values of latent variables are specified. The NML code-length for $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}$ is defined as the negative logarithm of the normalized maximum likelihood as follows:

$$-\log \frac{p\left(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}; \hat{\boldsymbol{\Theta}}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}); K, M\right)}{\int \mathrm{d}\boldsymbol{X}' \int \mathrm{d}\boldsymbol{Z}' \int \mathrm{d}\boldsymbol{S}' p\left(\boldsymbol{X}', \boldsymbol{Z}', \boldsymbol{S}'; \hat{\boldsymbol{\Theta}}(\boldsymbol{X}', \boldsymbol{Z}', \boldsymbol{S}'); K, M\right)}, \tag{2}$$

where $\hat{\boldsymbol{\Theta}}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S})$ is the maximum likelihood estimator of $\boldsymbol{\Theta}$ from $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}$. It is known [7] that it achieves the Shtarkov's minimax risk defined as follows:

$$\min_{\mathcal{L}} \max_{\mathbf{X},\mathbf{Z},\mathbf{S}} \left\{\mathcal{L}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}) - \min_{\boldsymbol{\Theta}}(-\log p\left(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}; \boldsymbol{\Theta}; K, M\right))\right\},$$

where the minimum is taken over all the prefix code-length function (i.e., the code-length function for uniquely decodable coding) for $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}$.

Note that the value of latent variables $\boldsymbol{Z}, \boldsymbol{S}$ are not observed, hence we estimate them from observable data $\boldsymbol{X}$ using the algorithm in Section 3.2, then plug them into the joint distribution, for which we use to calculate the NML code-length for $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}$. We call this process the NML coding with *latent variable completion* (LVC). Let $\hat{\boldsymbol{Z}}(\boldsymbol{X})$ and $\hat{\boldsymbol{S}}(\boldsymbol{X})$ be the estimates of $\boldsymbol{Z}$ and $\boldsymbol{S}$ from $\boldsymbol{X}$ and rewrite $\hat{\boldsymbol{\Theta}}(\boldsymbol{X}, \hat{\boldsymbol{Z}}(\boldsymbol{X}), \hat{\boldsymbol{S}}(\boldsymbol{X}))$ as $\hat{\boldsymbol{\Theta}}(\boldsymbol{X})$. We define the NML code-length for $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}$ with LVC as follows:

$$\begin{aligned}&\mathcal{L}\left(\boldsymbol{X}, \hat{\boldsymbol{S}}(\boldsymbol{X}), \hat{\boldsymbol{Z}}(\boldsymbol{X}); \hat{\boldsymbol{\Theta}}(\boldsymbol{X}); K, M\right)\\ &\stackrel{\mathrm{def}}{=} -\log \frac{p\left(\boldsymbol{X}, \hat{\boldsymbol{S}}(\boldsymbol{X}), \hat{\boldsymbol{Z}}(\boldsymbol{X}); \hat{\boldsymbol{\Theta}}(\boldsymbol{X}); K, M\right)}{\int \mathrm{d}\boldsymbol{X}' \int \mathrm{d}\boldsymbol{S}' \int \mathrm{d}\boldsymbol{Z}' p\left(\boldsymbol{X}', \boldsymbol{S}', \boldsymbol{Z}'; \hat{\boldsymbol{\Theta}}'; K, M\right)}.\end{aligned} \tag{3}$$

Here $\hat{\boldsymbol{\Theta}}'$ denotes the estimate of $\boldsymbol{\Theta}$ from $\boldsymbol{X}', \boldsymbol{S}', \boldsymbol{Z}'$. Note that the denominator in the right-hand side is analytically intractable. The following theorem yields an asymptotically intractable form of the NML code-length with LVC.

Theorem 1: Assume that each $\theta_{\omega k}^{(m)}$ is restricted to be in $[0, \theta_{\max}]$ for $\theta_{\max} < \infty$. The NML code-length (3) with LVC is asymptotically expanded within error $o(1)$ ($\lim_{n\to\infty} o(1) = 0$) as follows:

$$\begin{aligned}&\mathcal{L}\left(\boldsymbol{X}, \hat{\boldsymbol{S}}(\boldsymbol{X}), \hat{\boldsymbol{Z}}(\boldsymbol{X}); \hat{\boldsymbol{\Theta}}(\boldsymbol{X}); K, M\right)\\ &= -\log p\left(\boldsymbol{X}, \hat{\boldsymbol{S}}(\boldsymbol{X}), \hat{\boldsymbol{Z}}(\boldsymbol{X}); \hat{\boldsymbol{\Theta}}(\boldsymbol{X}); K, M\right)\\ &\quad + \frac{\Omega KM}{2} \log \frac{T\theta_{\max}}{\pi} + K \log R(\Omega, T, M),\end{aligned} \tag{4}$$

where

$$\begin{aligned}&R(\Omega, T, M)\\ &\stackrel{\mathrm{def}}{=} \int \prod_{t'=0}^{T-1} \mathrm{d}z_{0t'} \mathrm{Gamma}\left(z_{0t'}; a, b\right) \prod_{m=0}^{M-1} \left(\sum_{t=0}^{T-1} z_{0(t-m)}\right)^{\frac{\Omega}{2}}.\end{aligned}$$

The proof will be given in the full version.

We select the best number of bases so that the code-length(4) is minimized. Note that $R(\Omega, T, M)$ is dependent on Ω, T, M but independent of K. It can be computed with the Monte Carlo method in advance (note: Ω, T, M are given in advance), generating $\left\{z_{00}, z_{01}, \ldots, z_{0(T-1)}\right\} \in \mathbb{R}^T$. Theorem

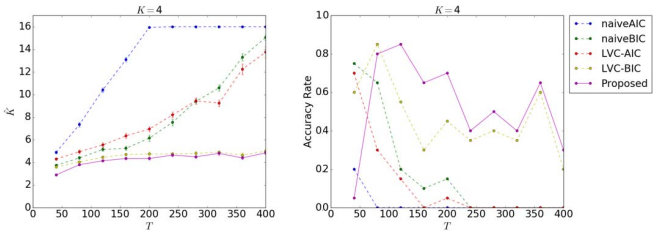

Fig. 3. Left: the number of bases estimated in the true structure estimation task, right: accuracy rates in the true structure estimation task

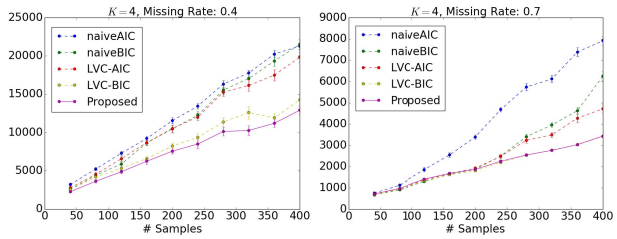

Fig. 4. The generalized Kullback Leibler divergence between the completed data matrix and the original data matrix without missing values. missing rate: 0.4 (left), 0.7 (right)

1 states that we can compute the NML code-length with LVC running the Monte Carlo simulation once in T dimensional space, while computing it straightforwardly in accordance with Equation (3) requires the Monte Carlo simulation in $\Omega T+KT$ dimensional space for all candidates of K.

V. Experiment

We show experimental results on structure selection for CNMF through synthetic data and real data. [1]

A. Synthetic Data Set

a) Data Set: We generated synthetic data sets as follows:

- $\theta_{\omega km} \sim \text{Uniform}\left(\theta_{\omega,k,m}; 0, 10\right)$
- $z_{kt} \sim \text{Gamma}\left(z_{k,t}; 2, 2\right)$
- $x_{\omega t} \sim \text{Poisson}\left(x_{\omega t}; \sum_{k=0}^{K-1}\sum_{m=0}^{M-1}\theta_{\omega km} z_{k(t-m)}\right)$,

for $\omega = 0, 1, \ldots, \Omega-1$, $k = 0, 1, \ldots, K-1$, $t = 0, 1, \ldots, T-1$, and $m = 0, 1, \ldots, M-1$, where $\Omega = 12$ and $M = 3$, and $K = 4$ and $T = 40, 80, \ldots, 400$.

We evaluated our method's performance from the two aspects; *true model estimation* and *missing data completion*.

b) Methods for Comparison: We employed for comparison the following conventional statistical criteria: naiveAIC, AIC applied naively by interpreting both of $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$ as parameters, which is given by $-\ln p\left(\boldsymbol{X}|\hat{\boldsymbol{Z}};\hat{\boldsymbol{\Theta}}\right) + (\Omega KM + KM)$, naiveBIC, BIC applied naively by interpreting both of $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$ as parameters, which is given by $-\ln p\left(\boldsymbol{X}|\hat{\boldsymbol{Z}};\hat{\boldsymbol{\Theta}}\right) + \frac{1}{2}(\Omega KM + KM)\ln T$, LVC-AIC: AIC applied with LVC, which is given by $-\ln p\left(\boldsymbol{X}, \hat{\boldsymbol{S}}(\boldsymbol{X}), \hat{\boldsymbol{Z}}(\boldsymbol{X}); \hat{\boldsymbol{\Theta}}; K, M\right) + \Omega KM$, and LVC-BIC: BIC applied with LVC, which is given by $-\ln p\left(\boldsymbol{X}, \hat{\boldsymbol{S}}(\boldsymbol{X}), \hat{\boldsymbol{Z}}(\boldsymbol{X}); \hat{\boldsymbol{\Theta}}; K, M\right) + \frac{1}{2}(\Omega KM)\ln T$ Note that naiveAIC and naiveBIC are *not* designed for model selection of irregular models. We did not consider the cross validation (CV) for comparison. This is because CV requires much more computation time than model selection criteria-based methods. In particular, learning CNMF takes much larger computational time than the conventional NMF, hence CV is not realistic for CNMF structure selection.

c) Results: Let us first show results on experiment of true model estimation. We took data matrices without missing as inputs. We then evaluated how well any method estimated the true number of bases, in terms of *accuracy* $A = \frac{1}{I}\sum_{i=0}^{I-1}\delta_{K_{\text{true}},\hat{K}^{(i)}}$ where I denotes the total number of trials, K_{true} denotes the true number of bases for the model generating synthetic data and $\hat{K}$ denotes the number of bases estimated by the respective methods. We set $I = 20$. The accuracy A shows the mean rate of how exactly the respective method was able to detect the true number of bases.

Fig. 3-left is a graph of the mean values of the estimated numbers of bases versus the temporal length T. Our method was able to detect the true numbers of bases more accurately than the competitive ones. Fig. 3-right shows accuracy rates for the respective methods. When the temporal length T was long enough, our proposed method showed much better performance than the competitive ones. This is because all of the competitive methods tended to overestimate the true numbers of bases as T increased. It also implies that when T was small, the competitive methods temporally got higher accuracy than our method because of their overestimation. We may see that our method also tended to slightly overestimate the true numbers of bases as T became sufficiently large. This is because the dimension of latent variables increased as T increased, hence the accuracy of estimating those variables as well as parameters became lower, which affected the accuracy of estimating the numbers of bases.

Next let us show results on experiment of missing data completion. We took data matrix with missing values as inputs. We then evaluated how well any method completed missing data in terms of the generalized Kullback-Leibler (KL) divergence as in Section II between the completed data matrix and the original data matrix without missing values. We conducted this test 20 times for each setting. The KL divergence was averaged over 20 trials.

Fig. 4 shows how well the respective methods completed missing value. In each graph, the horizontal axis shows the sample size while the vertical axis shows the generalized KL divergence. We observe that our method outperformed all of the competitive ones.

Through the experiments, we see that the proposed method worked better than the others both in the rank estimation and missing data completion scenarios. The superiority of the proposed method was more significant in the rank estimation scenario than in the missing data completion scenario.

[1] Python3 codes are available in https://github.com/atsushi-suzuki/cnmf_mdl

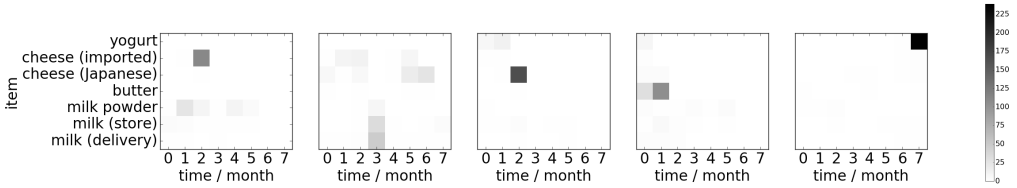

Fig. 5. Time sequential features in price fluctuation of dairy products in Japan extracted by CNMF

B. Price Fluctuation Data

Finally, we applied CNMF to price fluctuation data of dairy products in Japan. This is the first application of CNMF into the data sets other than acoustic data, as far as the authors know. We got the data from the statistic bureau, the ministry of internal affairs and communications of Japan[2] We used the price data of milk (delivered), milk (at a store), milk powder, butter, cheese (Japanese), cheese (imported) and yogurt, from January 2001 to March 2014. Note that we did not use the data from April 2014 to exclude the effect of the increase of Japan's consumption tax in April 2014. We first normalized the data by dividing them by the mean of each item's price. Then we obtained the absolute values of the difference between the target month's price and the last month's one, and multiplied 1000 by it. We applied CNMF to the preprocessed data. We set $b = 1.00001$, $a = 2.0$ and $M = 8$.

Fig. 5 shows 5 time sequential features extracted by CNMF. The darker the color is, the larger fluctuation it means.

The first base extracted by CNMF as in Fig. 5 shows the existence of the fluctuation of milk powder's price as a symptom of that of imported cheese's price. The second base shows a simultaneous fluctuation pattern among milk and milk powder and the existence of the fluctuation of cheese's price as a symptom of the simultaneous fluctuation. The third base shows the fluctuation of imported butter's price and the existence of the fluctuation of cheese's price as its symptom. The fourth and fifth bases show that the prices of butter and yogurt tended to fluctuate independently of other products.

Each of the 5 features represents a clearly distinct time sequential pattern of price fluctuation. Through the time sequential patterns, we may find chain reaction of price fluctuation among different products.

VI. Conclusion

This paper has proposed a novel methodology for selecting best structures of CNMF. Our structure selection method has been realized by 1) introducing a probabilistic structure with latent variables into CNMF, 2) designing an efficient algorithm for estimating both parameters and latent variables using the auxiliary function method, and 3) deriving a criterion for selecting the number of bases using the normalized maximum likelihood coding with latent variable completion under the minimum description length principle. Specifically, by 3), we are able to overcome the difficulty on the irregularity of the probabilistic modeling of CNMF. We have empirically demonstrated the effectiveness of our method using the synthetic and real data sets. We have further emphasized the usefulness of CNMF with our method in the scenario of data mining.

[2]Available in http://www.stat.go.jp/data/kouri/doukou/3.htm

Acknowledgment

This work was pentially supported by JST-CREST.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, Dec 1974.

[2] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009, 2009.

[3] S. Hirai and K. Yamanishi. Efficient computation of normalized maximum likelihood codes for gaussian mixture models with its applications to clustering. *IEEE Transactions on Information Theory*, 59(11):7718–7727, 2013.

[4] Y. Ito, S. Oeda, and K. Yamanishi. Selecting ranks and detecting their changes for non-negative matrix factorization using normalized maximum likelihood coding. In *SDM*, 2016.

[5] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama. Complex nmf: A new sparse representation for acoustic signals. In *ICASSP*, pages 3437–3440. IEEE, 2009.

[6] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001.

[7] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.

[8] J. Rissanen. *Optimal Estmation of Parameters*. Cambridge, 2012.

[9] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[10] Y. M. Shtar'kov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.

[11] P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Independent Component Analysis and Blind Signal Separation*, pages 494–499. Springer, 2004.

[12] T. Virtanen, A. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *ICASSP*, pages 1825–1828, March 2008.

[13] C. S. Wallace and D. M. Boulton. An information measure for classification. *The Computer Journal*, 11(2):185–194, 1968.

[14] S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11:3571–3594, 2010.

[15] S. Watanabe. A widely applicable bayesian information criterion. *The Journal of Machine Learning Research*, 14(1):867–897, 2013.

[16] K. Yamanishi. A learning criterion for stochastic rules. *Machine Learning*, 9:165–203, 1992.